基于 Bi-LSTM 和 CRF 的中文命名实体识别方法实现

姓名：焦新宇　班级：生研 2302　学号：2023201249

一、研究背景

1.1 研究背景与意义

命名实体识别（NER，Named Entity Recognition）是 NLP 众多高层应用通用的基础任务之一，其主要目标是从文本中识别和分类具有特定意义的实体。这些实体通常包括人物、组织、地点、日期、时间、货币、百分比等具体的命名实体。在诸如信息检索、机器翻译、问答系统等众多 NLP 技术中担任着非常重要的角色。为了解决这些挑战，研究者和从业者使用了各种技术，包括基于规则的方法、机器学习方法（如条件随机场、序列标注模型）以及最近流行的深度学习方法。

1.2 相关研究现状

NER 初期的研究中，因为可供研究的数据规模较小，采用的方法主要是基于人工规则的系统来识别实体，这些系统中的人工规则是由语言学知识丰富的专家们对命名实体的语法词法构成、上下文搭配及用词规律等进行分析后研究制定而成。

中期的研究中基于概率与统计的方法成为研究人员使用的主流方法，利用好统计学和概率学知识针对具体任务设计模型，然后从大量标注好的数据中进行监督学习后获得一个训练好的模型，最后使用训练好的模型对句子中的命名实体进行识别。常用的统计学模型有隐马尔可夫模型、条件随机场、支持向量机等。

近年来，基于深度学习的 NER 方法通过构建人工神经网络对非线性过程进行建模，可以在不同的数据集上自动学习抽象特征，取得了与统计学模型相当的效果，甚至大部分情况超过了统计学模型。而且深度学习方法对人工特征的依赖程度更低，跨领域适应性也比较强。

1.3 本文研究内容

实体识别（NER）的具体目标是从给定的文本中识别和分类出具有特定意义的实体。这些实体可以是文本中的任何命名实体，通常包括以下几类：1.人物：例如，"江小白"；2.组织：例如，"人民代表大会"；3.地点：例如，"北京"；4.日期：例如，"2023 年"。

NER 的目标不仅仅是识别这些实体，还要对它们进行分类，即确定实体属于上述哪一类。因此，NER 的输出通常是一系列实体标签，每个标签都对应文本中的一个实体，同时标明了实体的类型。在本任务中，主要通过性能、模型特性等目标深入了解深度学习方法的 NER 模型在不同方面的性能表现，为实际应用中的选择提供有价值的参考。

二、相关理论模型

2.1 条件随机场

中文命名实体识别任务中常采用 BIO 标注法。命名实体识别可以看作序列标注任务来处理。在序列标注中常使用线性链条件随机场。对于一个由 n 个字符构成的句子序列 X=(x1,X2..,.n)，标注序列 Y=(y1y2..yn)，令每个字符的标注标签 y；都在预设定的标签中任选其一。当为每个字符 x;都标注了标签 y；后,此时 Y 就形成了一个随机场。进一步的,当为标签的选择方式加一个限制条件：标签 yi 只与其上一个标签 yi-1 有关，此时随机场称作马尔可夫随机场。对于马尔可夫随机场中的两个参数，输入序列 K、标注序列 Y，X 是已给定的文字序列，此时的 Y 是在给定 X 的情况下的序列标注输出，于是 P(Y|X)便构成一个条件随机场。CRF 需要通过定义合适的特征特征函数（通常使用实值函数）作为特征函数来表示转移特征函数和状态特征函数。

2.2 长短期记忆网络

LSTM 网络是基于 RNN 改进的基本网络，其在 RNN 的神经单元中设置了各种门控记忆单元,用来控制当前时刻的输入以及上一时刻的输出。LSTM 单元在训练时，当前时刻的单元可以通过遗忘门、输出门以及输入门来控制上一个时刻的输出、当前时刻的输

入以及贯穿整个网络所截止到当前时刻的状态。LSTM 通过三个门控单元，使得 LSTM 网络具备选择记忆的功能。

2.3 Bi-LSTM-CRF 模型

Bi-LSTM-CRF 模型是将条件随机场 CRF 和双向长短记忆网络(Bidirectional long-term and short-term memory，Bi-LSTM)结合起来的深度学习命名实体识别模型，模型的结构一共可分为 3 层，嵌入层、Bi-LSTM 层、CRF 层，模型的整体结构如图 1 所示：

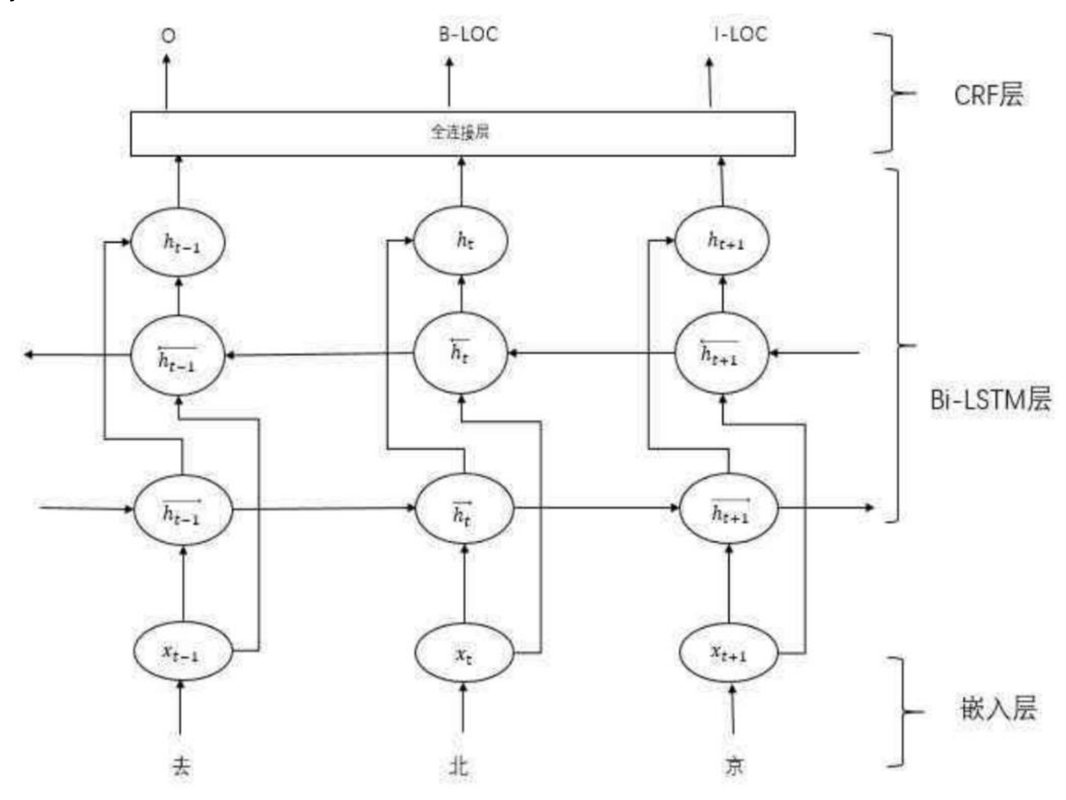


图 1. Bi-LSTM-CRF 模型

嵌入层："向量化"，主要作用是将高维稀疏向量转化为稠密向量，从而方便下游模型处理。

Bi-LSTM 层：LSTM 是 RNN 的一种，由于其设计的特点，非常适合用于对时序数据的建模， 如文本数据。Bi-LSTM 是由前向 LSTM 与后向 LSTM 组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。利用 LSTM 对句子进行建模还存在一个问题：无法编码从后到前的信息。通过 Bi-LSTM 可以更好的捕捉双向的语义依赖。

CRF 层：主要作用是对 Bi-LSTM 层的输出进行规范，避免出现一些不合理的序列。在 Bi-LSTM 网络生成素有的概率转移向量后，概率转移向量和标签就组成了一个条件随机场。当模型以减少目标损失函数为优化目标训练完毕后，CRF 层通过维特比算法从整个条件转移矩阵中找到一条最合理的序列作为预测标签，维特比算法是一个动态规划(Dynamic Programming，DP)算法。

优化算法：当在训练神经网络时，会出现这样的情况：模型用训练数据测试时有较高的准确率，而用测试数据测试时，准确率却大打折扣。这种情况称之为过拟合，过拟合在神经网络的训练数据不足而网络的参数又太多的时候更加明显。原则上认为，出现过拟合的模型是没有参考价值的，所以训练中要避免得到过拟合的模型。丢弃法(dropout)，用来解决过拟合问题。在神经网络迭代训练时，使用 dropout 让每个神经元有一定的概率停止工作。dropout 的工作原理如图 2 所示。

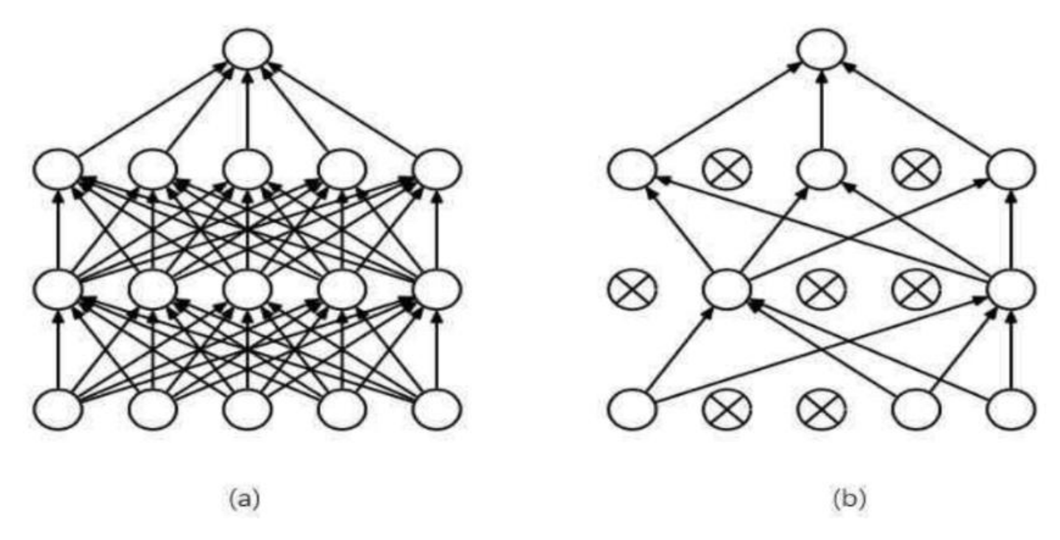


图 2. (a)使用 dropout 前，(b)使用 dropout 后

应用 dropout 之后的网络在每次的训练中，有一部分的神经元在训练时会被暂时忽略掉，这些神经元不会参与该次训练，而此时的网络只会更新那些未被忽略的神经元的参数。

三、Bi-LSTM-CRF 模型实现中文命名实体识别

3.1 实验数据预处理

3.1.1 数据集

MSRA 中文新闻数据集，该数据集是 Microsoft Research Asia 推出的关于中文命名实体识别的数据集，其中主要包括：地名、机构名和人名，采用的标签策略是 BIO，训练集含有 42000 个句子，3.4k+个地名（LOC），1.9k+个机构名（ORG），1.6k+个人名（PER）。验证集和测试集的大小大约是训练集的十分之一。

3.1.2 数据处理

1. 读取语料和标签信息，建立词汇和标签的映射词典，以下为标签及其映射示例。
   {'O': 0, 'B-ORG': 1, 'I-PER': 2, 'B-PER': 3, 'I-LOC': 4, 'I-ORG': 5, 'B-LOC': 6}
2. 分别加载训练集、验证集和测试集。选用 BertTokenizer 将纯文本转换为编码，该过程不涉及将词转换成为词向量，仅仅是对纯文本进行分词，并且添加[MASK]、[SEP]、[CLS]标记，然后将这些词转换为字典索引。

经过处理的数据包含三个部分：

- input_ids:词汇转化为编码的矩阵，使用 pad=0 填充。
- label_ids:对应标签转化为编码的矩阵，使用 pad=0 填充。
- mask：输入序列中有效词汇的位置标记为 1，其余为 0。

3. 将数据集进行封装为 Dataset 形式。

3.2 评估方法

- Precision：精确率又叫查准率，表示预测结果为正例的样本中实际为正样本的比例。
- Recall：召回率又被称为查全率，表示预测结果为正样本中实际正样本数量占全样本中正样本的比例。
- F1_score：精确率和召回率的一个加权平均。

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- Accuracy：准确率表示分类正确的样本占总样本个数的比例。

- 多分类 macro、micro、weighted：在多分类任务中，每一个类别都有一组 precision、recall、F1 分数，需要平衡各类别的分数得到全局的评价指标
  - macro：宏观，单个类别计算完之后取平均，不考虑类别不平衡问题。
  - micro：微观，所有样本做一个整体，分母为所有样本个数。
  - weighted：按真实值中各类别个数取权重。

3.3 实验参数设置

训练按照四步走策略开展：

1. 计算当前网络权重下的正向输出结果和实际结果的损失；
2. 梯度置为 0，防止梯度累计；
3. 反向传播；
4. 梯度下降，执行参数更新;
5. 采用一轮训练，一轮测试的方式。

参数配置：

- embed_size=100：嵌入层大小
- hidden_size=128：隐藏层大小
- num_layer=1：LSTM 的层数
- lr=0.001：学习率
- optimizer=Adam：优化器
- loss=CrossEntropyLoss：交叉熵损失函数

3.4 实验结果分析

3.4.1 学习率

在学习率为 0.001 时，曲线变化更加平滑，能得到效果更好的模型。

3.4.2 Dropout 值

在初始的实验中出现过拟合的情况，在不同 dropout 值下进行了实验。

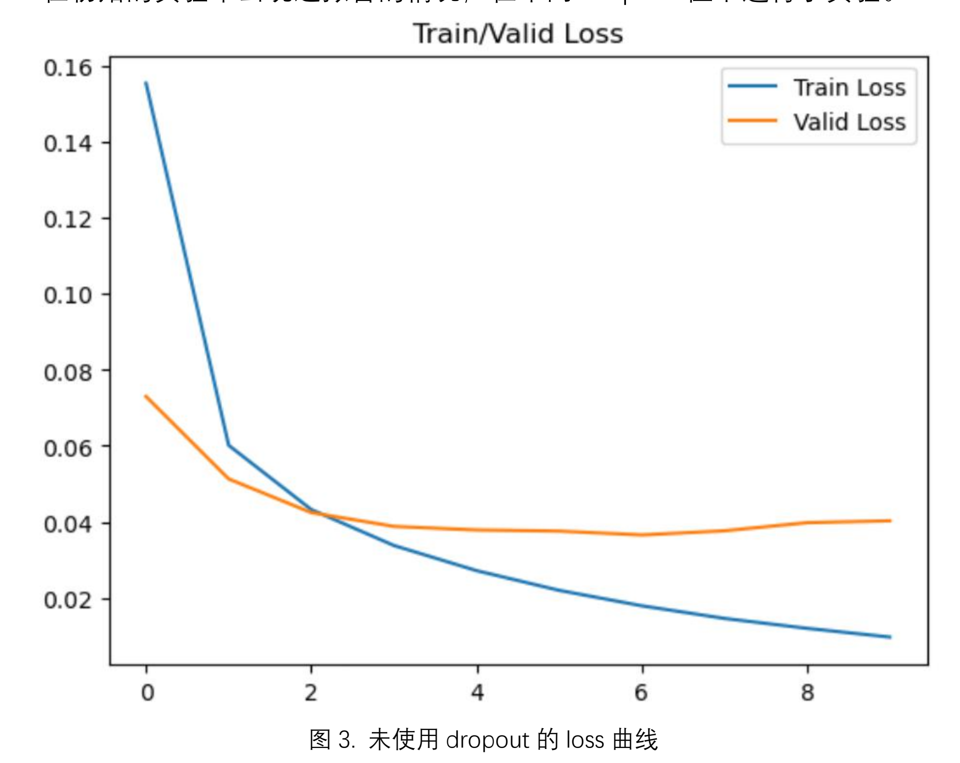


图 3. 未使用 dropout 的 loss 曲线

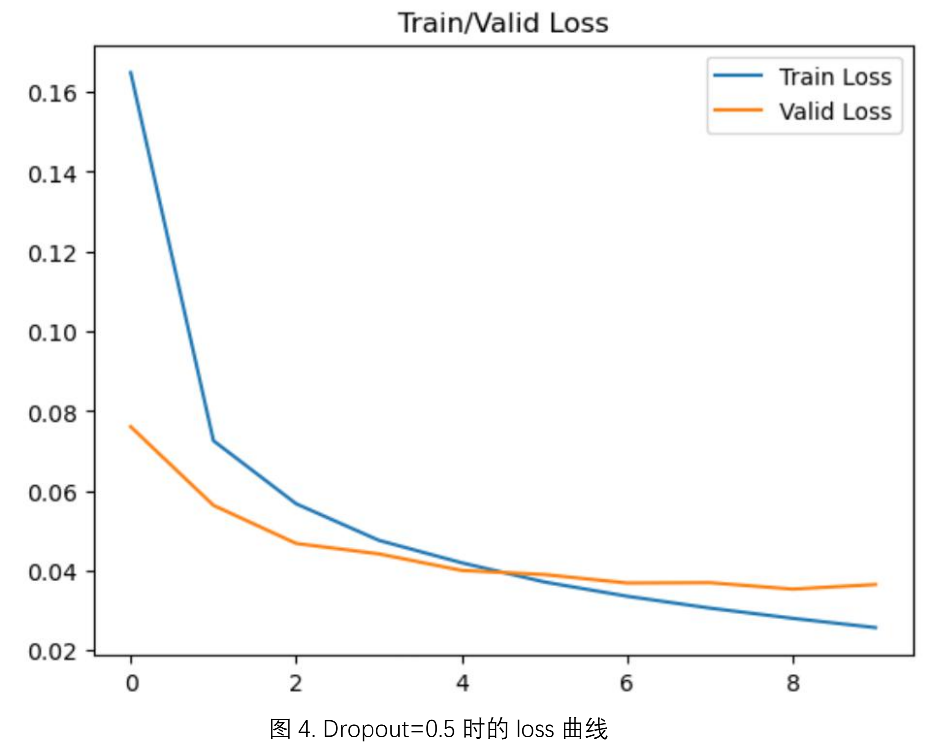


图 4. Dropout=0.5 时的 loss 曲线

从图中可以看出，使用 dropout 效果比无 dropout 效果好，本文选用 dropout=0.5 开展实验。此外，尽管使用 dropout 减少了过拟合的情况，但模型仍然过拟合。因此在任务中使用 L1 正则化参数进行了实验，但发现尽管模型拟合很好，但预测精确率等参数显著降低，因此放弃使用 L1 参数。

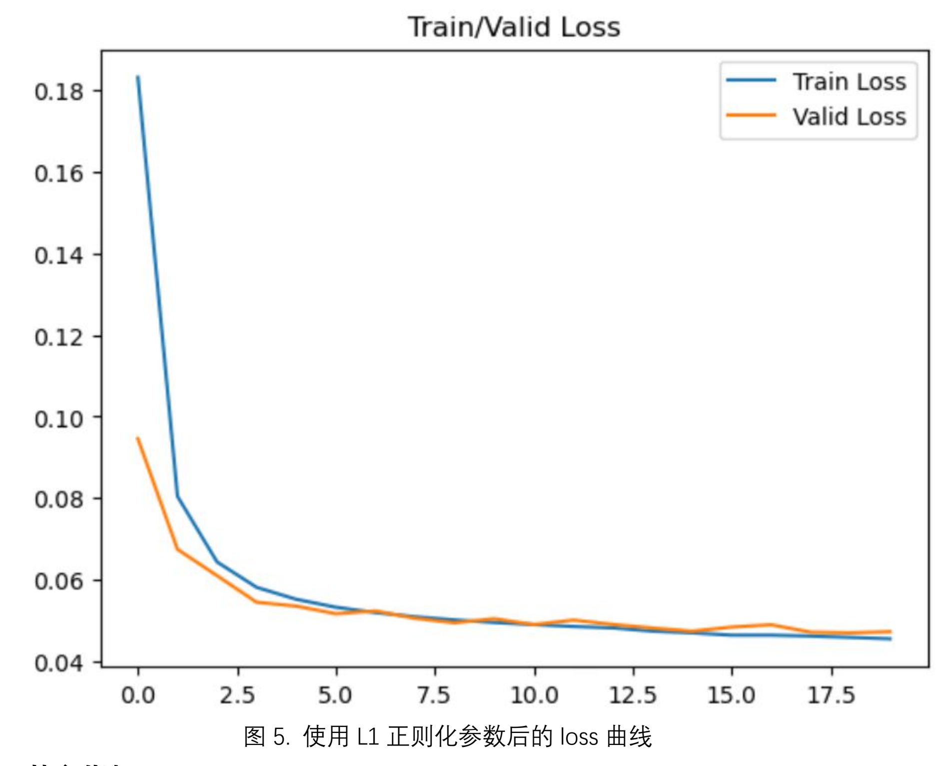


图 5. 使用 L1 正则化参数后的 loss 曲线

3.4.3 其它指标

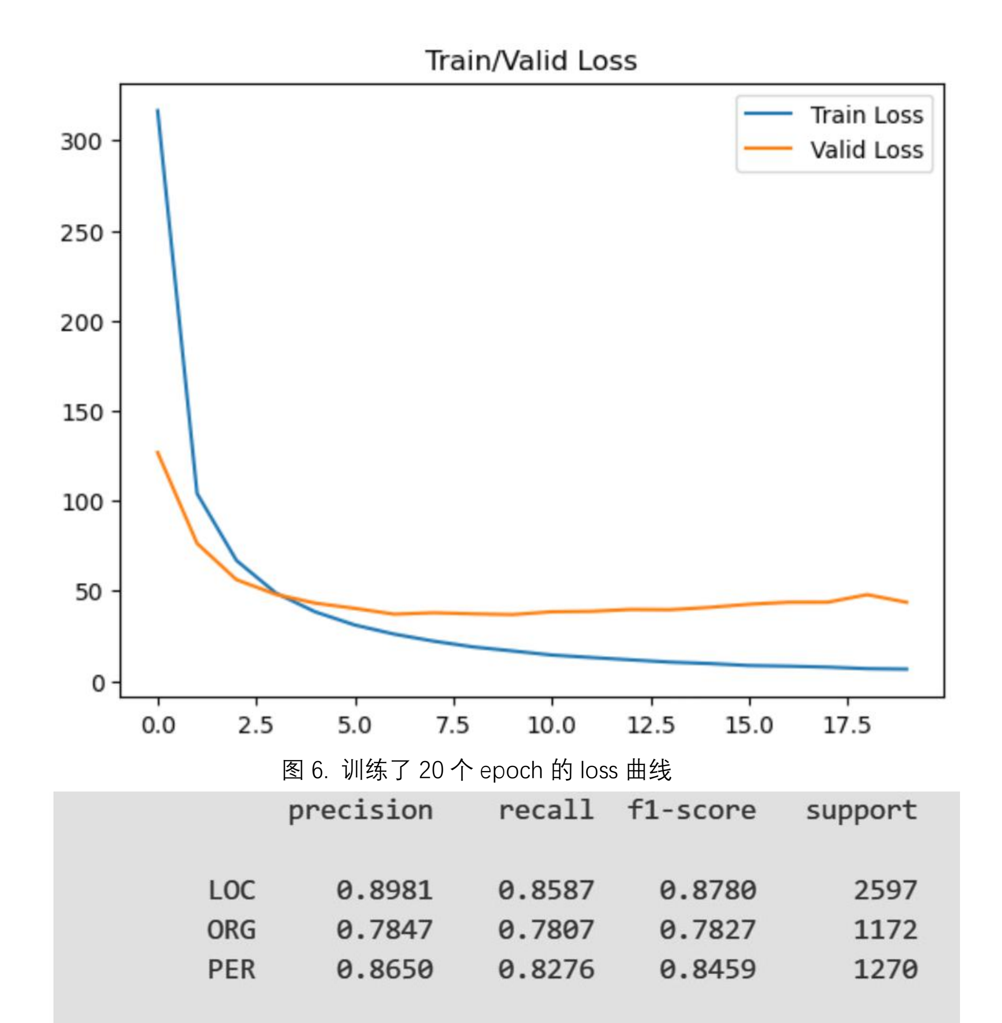


图 6. 训练了 20 个 epoch 的 loss 曲线

	precision	recall	f1-score	support
LOC	0.8981	0.8587	0.8780	2597
ORG	0.7847	0.7807	0.7827	1172
PER	0.8650	0.8276	0.8459	1270
micro avg	0.8627	0.8327	0.8474	5039
macro avg	0.8493	0.8223	0.8355	5039
weighted avg	0.8634	0.8327	0.8477	5039

图 7. 精确率、召回率、F1 分数等指标

```
高举爱国主义和社会主义两面旗帜，团结全体成员以及所联系的归侨、侨眷，发扬爱国革命的光荣传统，为统一祖国、
test data:
        LOC: 中华
        ORG: 致公党中央
        ORG: 中共中央
        ORG: 国务院

pred result:
        LOC: 中华
        ORG: 致公党中央
        ORG: 中共中央
        ORG: 国务院
```

图 8. 预测示例

3.4.4 实验结果分析

评价指标分析结果表明，ORG 的识别率低，可能是复合词占比较高，组织名中包含人名或地名，造成识别错误。如 “美中关系全国委员会” -> “美中” 和 “全国委员会”。且查看 CRF 状态转移特征发现：存在 I-LOC->B-ORG 和 I-PER->B-ORG 的状态转移，表明模型学习到的数据集中包括这两个特性，组织名中包含人名和地名造成识别错误的可能确实存在。

```
Top likely transitions:
I-ORG  -> I-ORG   6.266624
B-PER  -> I-PER   3.919506
I-PER  -> I-PER   3.626846
O      -> O       3.560464
B-ORG  -> I-ORG   3.428959
I-LOC  -> I-LOC   2.929716
B-LOC  -> I-LOC   2.910829
O      -> B-ORG   1.491918
O      -> B-PER   0.588304
O      -> B-LOC   0.410751
I-LOC  -> O       -0.077332
I-LOC  -> B-LOC   -0.078167
I-PER  -> O       -0.092089
I-ORG  -> O       -0.248649
I-PER  -> B-PER   -0.745519
B-LOC  -> B-LOC   -0.927507
I-LOC  -> B-ORG   -1.144386
I-PER  -> B-ORG   -1.279218
I-ORG  -> B-ORG   -1.338060
B-LOC  -> O       -1.516935
```

图 9. CRF 状态转移特征

任务中存在过拟合且准确率较低的原因可能是使用的数据集偏小，对于需要大量样本的深度学习而言，样本略少。从场景来说，如果需要识别的任务不需要太依赖长久的信息，此时 RNN 等模型只会增加额外的复杂度。

四、结论

随着通信技术和互联网的发展的突飞猛进，互联网的数据规模一直在不断的扩张，这些数据虽然蕴含了大量可以被利用的信息，但是却常常因其非结构化的形式而很难被直接利用。信息抽取作为文本信号处理的重要研究领域，其主要研究目标便是通过计算机从非结构化的数据中自动提取结构化的信息。命名实体因为是这些结构化信息中的重要组成部分。能准确地对其进行识别将会直接影响信息抽取技术的准确性。此外，对中文语料的命名实体识别研究主要在集中在国内，而国内的技术起步相较于一些发达国家要晚，技术也相对陈旧，所以对命名实体识别尤其是对中文命名实体识别的研究具有非常重要的价值与意义。

本任务使用的中文命名实体识别模型是基于深度学习的 Bi-LSTM-CRF 模型，最终实现的 f 值为 84%。但是由于数据集等限制，本任务仍然有一些值得进一步研究的方向。未来优化的方向在于扩大数据集、根据实际数据做出模型调整或者优化模型架构。